時系列の状態空間モデル・時系列の機械学習

今日の目標

- 時系列モデルの状態空間モデルを説明できる
- 時系列予測ライブラリProphetが使える

状態空間モデル

時系列の隠れた変数

琵琶湖の漁の毎年の漁獲量は時系列. 過去の漁獲量(kg)のデータを見て, $\mathrm{AR}(2)$っぽい, とか, ϕ_1, ϕ_2, σ^2 はいくつ? とか推定することはできる. しか〜し.

- 実は背後に, 琵琶湖の生息するナマズの総数とか, モロコの総数とか, もっと言えば, 水温とか水草の量とか, 観測できないより本質的な時系列があるのでは?
    - 被食者捕食者の増減の話とかやったよね?/やります.微分方程式, 数理モデル
- ホワイトノイズϵ_tっていうけど, 今年はたまたま稚魚が育たなかったっていうノイズと, 今年はコロナで漁船の整備ができなかった(?)っていうノイズは別なのでは?

システムと観測を分けて, もっと精密に考えたい!

状態空間モデル

状態空間モデルの簡単な場合

- 観測できない量
    - $\mathbf{X}_t$: 状態ベクトル(2次元)
    - $\mathbf{v}_t$: システムノイズベクトル(2次元)
    - w_t: 観測ノイズ(1次元)
    - $\mathbf{v}_t$,w_t は互いに独立なホワイトノイズ
- 観測できる量
    - Y_t: 観測値(1次元)
- 定数
    - F, G: 2×2行列
    - H: 1×2行列

$$\text{システムモデル}\mathbf{X}_t = F\mathbf{X}_{t-1} + G\mathbf{v}_t$$
$$\text{観測モデル}Y_t = H\mathbf{X}_t + w_t$$

状態空間= $\mathbf{X}_t$ の観測できない空間

状態ベクトルは, $\mathbf{X}_t = F\mathbf{X}_{t-1} + G\mathbf{v}_t = F(F\mathbf{X}_{t-2} + G\mathbf{v}_{t-1}) + G\mathbf{v}_{t-1} = \cdots = F^t\mathbf{X}_0 + \sum_{t'=0}^{t} F^{t-t'}G\mathbf{v}_{t'}$ のように漸化式で計算される.

観測値 Y_t は漸化式で結ばれておらず, 各時刻ごとに $Y_t = H\mathbf{X}_t + w_t$ で$\mathbf{X}_t$ から計算される.

例

例:琵琶湖の漁獲量

$$F=\begin{pmatrix}1.01 & 0.3\\ -0.5 & 0.98\end{pmatrix},\mathbf{X}_t=\begin{pmatrix}\text{なまずの個体数}\\ \text{モロコの個体数}\end{pmatrix},G=\begin{pmatrix}1 & 2\\ 100 & 200\end{pmatrix},\mathbf{v}_t=\begin{pmatrix}\text{正規分布(水草の量)}\\ \text{正規分布(水温)}\end{pmatrix}$$

$$Y_t=\text{漁獲量(kg)},H=\begin{pmatrix}1000 & 10\end{pmatrix},\mathbf{X}_t=\begin{pmatrix}\text{なまずの個体数}\\ \text{モロコの個体数}\end{pmatrix},w_t=\text{正規分布(漁船の数)}$$

例:AR(2)

$$x_t=\phi_1x_{t-1}+\phi_2x_{t-2}+\epsilon_t.$$

は,

$$\begin{pmatrix}x_t\\ x_{t-1}\end{pmatrix}=\begin{pmatrix}\phi_1 & \phi_2\\ 1 & 0\end{pmatrix}\begin{pmatrix}x_{t-1}\\ x_{t-2}\end{pmatrix}+\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\text{正規分布}\\ \text{正規分布}\end{pmatrix},$$
$$y_t=\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}x_t\\ x_{t-1}\end{pmatrix}+0$$

と表せる.

AR(3)は

$$\begin{aligned}x_t&=\phi_1x_{t-1} && +\phi_2x_{t-2}+\phi_3x_{t-3} && +\epsilon_t,\\ x_{t-1}&=1\cdot x_{t-1},\\ x_{t-2}&= && +1\cdot x_{t-2},\\ Y_t&=x_t && && +0\cdot\epsilon'_t\end{aligned}$$

のように3次元の状態空間で表せる.

例:ローカルレベルモデル

$$\begin{aligned}x_t&=x_{t-1}+\text{乱数},\\ y_t&=x_t+\text{別の乱数}\end{aligned}$$

ランダムウォークだが, 観測者によるウォーカーの位置の測定値 y_t は正確ではない. その前提で, 正確な位置 x_t を求めたい.

状態空間モデルの特徴

- $\mathrm{AR}(m)$ や親戚の MA, ARMA, SARIMA などのモデルを統一的に表せる.
- これでないと表せないモデルもある.
- 制御工学でよく扱われて, 解法が整備されている. フィルタ, 平滑化. カルマンフィルタ pykalman
- パラメタが多い
- HMM=Hidden Markov model 隠れマルコフモデル はこの離散バージョン

状態空間モデルを使う手続き

1. 与えられた y_t を見る
2. (原理から考えて)モデルを書く
3. F,G,Hが未知ならそれをパラメタ推定する

4. 与えられた学習データ $\{y_t\}_{t=0,\ldots,T}$ から $\mathbf{X}_t$ を状態推定(予測,フィルタ,平滑化)する.
5. テストデータ y_t と, $\mathbf{X}_t$ から求めた $\hat{y}_t$ を比較して, 性能を評価する

Python の関数 statsmodels.tsa.UnobservedComponents

状態空間モデルの扱いやすい場合1

- $\mathbf{v}_t, \mathbf{w}_t$ が正規分布にしたがうと仮定する. このとき, 正規分布の加法性から, $\mathbf{X}_t, Y_t$ などがすべて正規分布にしたがい, それらの母平均値, 母共分散の間の関係式をつかって解析できる.
- 制御工学でよく扱われて, 解法が整備されている. フィルタ, 平滑化. カルマンフィルタ pykalman

状態空間モデルの扱いやすい場合2:隠れマルコフモデル

- 「$\mathbf{X}_t$が離散型確率変数の場合」を考えたい. $X_t = 1, 2, \cdots, n$.
- システムモデルは「 $\mathbf{X}_t$ が確率的に時間発展する」ことを表すので, マルコフ連鎖で書くのが自然. Mは推移確率行列. 計算科学B, 確率モデル及び実習

$$
\begin{aligned}
P(X_t = 1) &= M_{11}P(X_{t-1} = 1) + M_{12}P(X_{t-1} = 2) + \cdots + M_{1n}P(X_{t-1} = n)\\
P(X_t = 2) &= M_{21}P(X_{t-1} = 1) + M_{22}P(X_{t-1} = 2) + \cdots + M_{2n}P(X_{t-1} = n)\\
&\vdots\\
P(X_t = n) &= M_{n1}P(X_{t-1} = 1) + M_{n2}P(X_{t-1} = 2) + \cdots + M_{nn}P(X_{t-1} = n)
\end{aligned}
$$

- 観測できない(隠れた)$X = 1, 2, \cdots, n$という状態があり, ある法則で時間発展する. 観測されるY_t の値はX から決まる.

時系列の機械学習

因果関係, ある時点以前のデータから以降のデータを予測することの需要が大きい, ことから次のように行われる

1. 観測データ x_tを 前半の $x_{\mathrm{train}t}$と, 後半の $x_{\mathrm{test}t}$とに分ける
2. $x_{\mathrm{train}t}$から, モデル, パラメタを推定する.
   - 誤差最小や尤度最大, などの原理がありうるが, パラメタの個数を変える変更(例$\mathrm{AR}(m)$のmの変更)の際は, AIC, BICなどの情報量規準で考えることになる.
3. 推定したモデル, パラメタから, $\hat{x}_t$ を作る.
4. $\hat{x}_t$ と $x_{\mathrm{train}t}$ との誤差を確認する. [statsmodelsでは predict() で求められることが多い. fittedvalues という名前になっていることがある] 初項から何個かは $\hat{x}_t$ が作れないことがある.
5. $\hat{x}_t$ と $x_{\mathrm{test}t}$ との誤差を確認する. [statsmodelsでは predict() で求められることが多い.]

ライブラリ pmdarima

SARIMA(p,d,q,P,D,Q,m)用のの関数. statsmodels の ar_model や arma_model とはまた別のフレイバー.

Notebook参照

ライブラリ Prophet

日ごとのデータを扱う ライブラリ. Facebook が開発するオープンソース. 祝日のデータなども総動員して自動化されている.

Notebook参照

その他の時系列の機械学習

- RNN=Recurrent Neural Network, LSTM
- 自然言語は, 順番がついた文字や単語の列であることから, 時系列解析と共通点がある
